Language Translation using Neural Machine Translation for Low-Resource Language Pairs on Hindi to Rajasthani Language

Synopsis of the Research Proposal
Submitted for the Registration in the

Ph.D. Program (Part-Time) in Computer Application

By

Yogender Kumar Saini

Under Supervision of

Dr. Rana Mukherji

Dean
IcfaiTech
(Faculty of Science and Technology)

THE ICFAI UNIVERSITY JAIPUR

Table of Content

Sr. No.	Title	Page No.
1	Proposed Summary	3
2	Introduction	4
3	Literature Review	6
4	Present Study	9
5	Methodology	15
6	Limitations	21
7	Work Plan of the Research	22
8	References	22

Proposed Summary / Abstract

This research work focuses on designing and building a system that can automatically translate text from Hindi to Rajasthani Language using Natural Language Processing (NLP) techniques. Hindi is one of the most widely spoken and officially recognized languages in India, and it has a strong digital presence. On the other hand, Rajasthani Language is a major regional language spoken mainly in Rajasthan and nearby areas, but it has very little digital content and is not yet fully supported by modern technology tools.

The main goal of this research is to bridge this gap by creating a translator that understands Hindi sentences and converts them correctly into Rajasthani Language. For this, we will collect Hindi-Rajasthani Language text data and prepare it for use by the system. This process will include cleaning the text, identifying word types (nouns, verbs, etc.), and analyzing sentence structures. We will use different NLP methods such as rule-based translation, statistical models, and machine learning to find the best way to translate while keeping the meaning and context intact.

Special care will be taken to handle local expressions, grammar differences, and word meanings that change with context. Since Rajasthani Language is a low-resource language (with fewer digital tools and datasets), this research will also help create useful resources like a Rajasthani Language dictionary and a bilingual dataset.

The final system will help people who speak Rajasthani Language to access digital content more easily, and it will support the preservation and promotion of the Rajasthani Language . This translator can also be used in mobile apps, educational platforms, and voice assistants in the future. Overall, this project will contribute to the growth of Indian language technology and promote regional language inclusion in the digital world.

1. Introduction

Language helps people to talk, share ideas, learn new things, and stay connected. In a country like India, where many languages are spoken, it is important to make communication easy for everyone. Hindi is one of the most common and official languages in India. It is used in schools, government work, newspapers, websites, and mobile apps[3]. A large amount of digital content is available in Hindi.

So, language is one of the most important tools for communication. It helps people talk, learn, and share ideas. In India, many different languages are spoken. Hindi is one of the most commonly used languages and is recognized officially in the country[2]. Because of this, there is a lot of content like books, websites, news articles, and mobile applications written in Hindi.

However, many people in India, especially in small towns and villages, speak regional languages. One such language is Rajasthani Language, which is mostly spoken in Rajasthan and nearby areas. Rajasthani Language is a beautiful and rich language, but it is not recognized as an official language and is rarely used in digital or written content[4]. As a result, Rajasthani Language-speaking people may face difficulties in understanding Hindi content on websites, government portals, and educational materials.

This research focuses on creating a Hindi to Rajasthani Language translation system using Natural Language Processing (NLP)techniques. NLP is a part of artificial intelligence (AI) that helps computers understand and work with human language. With NLP, machines can read, understand, and even translate languages[5].

To build a good translation system, the first step is to collect a parallel dataset. This means collecting a set of Hindi sentences along with their correct Rajasthani Language translations. This data will be cleaned and processed so that a computer can learn from it. We will study the grammar, word order, and structure of both Hindi and Rajasthani Language to find out how they are similar or different[8].

There are different methods to do machine translation. One way is rule-based translation, where we create language rules manually. Another method is statistical machine translation (SMT), which learns from large amounts of data. Recently, neural machine translation (NMT) has become very popular because it uses deep learning models like the Transformer[8]. These models work very well when large datasets are available, but they are hard to apply for low-resource languages like Rajasthani Language [19].

Our main goal is to create a special computer program that can automatically translate text from Hindi into Rajasthani Language.[10][11]. Think of it like a smart "Google Translate" but specifically designed for Rajasthani Language. Why is this important? Well, Rajasthani Language is a very old and beautiful language, spoken by many people here in Rajasthan. It's full of unique phrases, sayings, and a rich cultural history[12].

In contrast, Rajasthani Language is primarily spoken in western Rajasthan and adjacent regions by 45–50 million speakers, yet it remains largely undocumented and under-supported in digital technology [15]. Rajasthani Language exhibits significant lexical and structural variation across dialects, sharing roughly 50–65% similarity with Hindi, but with distinct phonetic correspondences (such as Hindi /s/ vs. Rajasthani Language /h/) and regional variation even within Rajasthan [11]. Despite its rich linguistic heritage, Rajasthani Language lacks institutional recognition and is seldom featured in modern NLP research or language technology initiatives.

Machine Translation Trends in India, India's MT research has evolved steadily, from rule-based frameworks to statistical systems and now neural translation models. First Rule-Based Machine Translation (RBMT) systems rely on manually crafted grammar rules and bilingual lexicons. Projects like Anuvaad (Hindi–Sanskrit) demonstrate that RBMT can yield high accuracy in closely related language pairs, especially with limited digital resources [15]. Second Statistical Machine Translation (SMT) grew popular with tools that learn translation patterns from bilingual corpora. CDAC's Mission Hindi project used SMT with suffix separation and pre-ordering to address structural divergence between English and Indian languages including Hindi [18]. Similarly,

A hybrid English–Hindi MT system integrated RBMT and SMT approaches, showing performance improvements in translation tasks despite low resources [19]. Third Neural

Machine Translation (NMT) using RNNs and Transformers has transformed MT quality and fluency. However, NMT's reliance on large-scale parallel training data limits its applicability for low-resource pairs like Hindi–Rajasthani Language. Researchers have explored complementary strategies such as multilingual and transfer learning, back-translation, and synthetic corpora augmentation to mitigate data scarcity [10].

2. Literature Review

The development of natural language translation systems in India has seen substantial progress in recent decades, especially in the area of Hindi and other regional language pairs. However, the Hindi-to-Rajasthani Language pair remains underexplored despite Rajasthani Language being spoken by over 45 million people, predominantly in Rajasthan. Rajasthani Language's absence in formal education, digital platforms, and official recognition poses unique challenges for NLP researchers [15].

Language diversity in India requires robust and inclusive NLP systems. Research shows that over 450 languages are spoken in the country, with Hindi playing a dominant role in communication, governance, education, and digital content creation [3][10]. This makes Hindi a suitable source language for translation, especially to less-resourced languages like Rajasthani Language. Despite Rajasthani Language's similarity to Hindi in structure and vocabulary, it lacks comprehensive documentation and digital tools for translation [4][11].

Initial machine translation systems in India were rule-based (RBMT), manually encoding grammar rules and lexicons. Anuvaad, a Hindi–Sanskrit MT system, demonstrated high effectiveness in closely related languages using RBMT approaches [15]. However, rule-based systems require extensive linguistic expertise and may not scale well to dialectally diverse languages like Rajasthani Language.

Statistical Machine Translation (SMT) gained popularity due to its ability to learn translation patterns from bilingual corpora. Projects such as Mission Hindi employed techniques like suffix separation and pre-ordering to manage linguistic divergences between English and Indian languages [18]. A hybrid English-Hindi system combining SMT and RBMT also showed promising improvements, particularly in low-resource scenarios [19].

With the rise of deep learning, Neural Machine Translation (NMT) has emerged as a dominant approach, especially using Transformer models [19]. While NMT systems achieve higher fluency and translation quality, they are heavily data-dependent and perform poorly with limited corpora. This is a major challenge for Hindi-Rajasthani Language translation, where no large parallel dataset currently exists [10].

Several strategies have been proposed to overcome data scarcity in low-resource settings. These include back-translation, data augmentation, and transfer learning from related language pairs. In Indian contexts, multilingual and character-level models have also been explored for improved performance [16][17]. These methods offer promising directions for Hindi-Rajasthani Language translation development.

The creation of a parallel dataset of Hindi-Rajasthani Language is the foundational step for any supervised MT system. dataset design involves selecting representative texts, aligning sentence pairs, and ensuring linguistic fidelity [8]. In the absence of prior Rajasthani Language corpora, researchers must rely on manual translation, community-sourced content, or mining parallel text from bilingual websites and literature.

Sociolinguistic studies highlight the endangered status of many regional Indian languages, including Rajasthani Language. Without technological support, such languages risk further marginalization. A translation system tailored to Rajasthani Language not only aids communication but also supports cultural preservation and digital inclusion [12][15].

Integrating NLP into regional language support is more than a technological challenge—it is a cultural and social imperative. A Hindi–Rajasthani Language translator would enhance access to information, education, and services for Rajasthani Language speakers. Furthermore, it contributes to the preservation of Rajasthan's cultural identity, deeply tied to the Rajasthani Language through oral traditions, folk literature, cuisine, and rituals [22].

Relevance to Hindi–Rajasthani Language Translation, There is no known academic work specifically on Hindi-Rajasthani Language machine translation. Yet, Rajasthani Language's lexical similarity and shared Indo-Aryan origin with Hindi make it amenable to techniques applied to related language pairs. SMT with suffix handling and pre-ordering helps address

divergence even in typologically close languages [8], while hybrid models combining linguistic rules and statistical learning provide robustness in low-data settings[4].

Rule-based systems like Anuvaad work effectively for Sanskrit-Hindi and can inspire grammatical mapping between Hindi and Rajasthani Language where explicit rules might compensate for sparse data [15]. Meanwhile, data augmentation methods using phrase-level back-translation and character-level encoding have improved Arabic and Hindi translations previously [16]. Techniques like multilingual low-resource representation using multilingual encoding (WX or BPE) also show performance gains in similar-language translation scenarios [17].

The successful development of a Hindi to Rajasthani Language translator carries profound cultural and social implications. It will significantly enhance digital literacy and access to information for the Rajasthani Language-speaking population, thereby fostering improved communication and understanding within the linguistically rich tapestry of Rajasthan and beyond [15]. By providing a dedicated and functional translation tool, this research directly contributes to the preservation and promotion of the Rajasthani Language itself.

This effort is crucial in mitigating the risk of language loss, a concern for many underserved languages, and ensures Rajasthani Language's continued vitality and relevance in the digital age. Furthermore, it supports the rich cultural heritage of Rajasthan, which is deeply intertwined with the Rajasthani Language language, evident in traditions such as the Ghoomar dance and the unique Rajasthani Language Bhojnalaya cuisine. This project becomes an act of cultural affirmation and preservation, ensuring that a significant linguistic identity can thrive in the digital sphere [20][22].

In conclusion, while Indian researchers have made commendable progress in NLP for major Indian languages, there remains a clear research gap in Rajasthani Language translation. This PhD research aims to bridge that gap by designing and implementing a Hindi-to-Rajasthani Language translation system using cutting-edge NLP techniques, tailored to the linguistic, technical, and sociocultural dimensions of this low-resource language pair.

3. Present Study

3.1 Problem Statement

Rajasthani Language, primarily spoken in Rajasthan and parts of Gujarat, Haryana, and Madhya Pradesh. Though widely spoken, Rajasthani Language lacks standard grammar, script, and digital resources. It is considered a dialect of Rajasthani and has not been officially recognized as a separate language, which has led to its exclusion from most language technology developments.

To the best of our knowledge, there is no prior work that systematically applies Neural Machine Translation (NMT) techniques to the Hindi–Rajasthani Language pair. This study represents the first comprehensive and structured attempt to build and evaluate an NMT-based translation system specifically for Hindi to Rajasthani Language, a low-resource and underrepresented language pair in existing research literature.

On the other hand, Hindi is one of the most spoken and studied languages in India. It enjoys government support, a standard script (Devanagari), rich grammar, and large-scale digitized corpora. It is included in most machine translation systems and has become a core language for various Natural Language Processing (NLP) applications.

One of the major challenges in developing a Neural Machine Translation (NMT) system for the Hindi–Rajasthani Language pair is the extreme scarcity of parallel corpora. Unlike high-resource language pairs, Hindi–Rajasthani Language lacks publicly available, large-scale aligned datasets required to train deep learning models effectively.

Transfer learning involves using a model that has already been trained on a large dataset for a high-resource language like Hindi to start the training for a low-resource language pair like Hindi-Rajasthani Language. Instead of training a new model from scratch, we're using a "pre-trained" model that has already learned general linguistic patterns, grammar, and vocabulary.

The field of Natural Language Processing (NLP), especially machine translation (MT), has witnessed transformative advancements, largely propelled by the availability of vast digital corpora and sophisticated computational models for high-resource language pairs (e.g.,

English-French, Chinese-English, even Hindi-English). However, a critical and pronounced research gap persists in the domain of low-resource languages, where such digital assets are scarce or non-existent. The specific and most significant research gap lies in the systematic, comprehensive, and academically rigorous investigation of Hindi-to-Rajasthani Language machine translation.

Despite Rajasthani Language's widespread use, there has been no major academic or technological study on creating a translation system between Hindi and Rajasthani Language. A review of linguistic resources, academic journals, and government projects reveals a complete absence of digital corpora, grammar rules, language models, or translation engines for this language pair. This shows a major gap in the field of regional language processing in India.

There is also no documentation of Rajasthani Language in computational linguistics or NLP literature. Most Indian language translation research focuses on high-resource languages or official languages recognized by the constitution. This has left regional dialects like Rajasthani Language out of the digital revolution.

Here's why translating between Hindi and Rajasthani Language using computers is so use full.

a) Missing Study Books

Imagine you want to teach a computer how to translate. The best way to do this is to show it lots and lots of examples. Think of it like a child learning a language by seeing words in English and then seeing the exact same words in Hindi.

For computers, these "study books" are called "parallel corpora." This means huge collections of text where every sentence in Hindi has its perfect, matching translation in Rajasthani Language, side-by-side.

The problem is, these "study books" for Hindi and Rajasthani Language simply don't exist in large enough numbers. It's like trying to teach a student without enough textbooks. We might find a few small examples here and there (like old stories or religious texts), but nothing big

enough for a computer to truly learn how to translate well. This is the biggest reason why Hindi-to-Rajasthani Language translation is so difficult right now.

b) Missing Dictionaries and Grammar Rules for Computers

Every language has its own words (vocabulary) and its own rules for how words change and how sentences are put together (grammar). Rajasthani Language isn't just Hindi with a different accent; it has unique words and its own way of forming sentences.

For computers to translate, we need to "know" these words and rules. we need: Digital dictionaries: Not just paper ones, but ones that computers can easily read and use. Grammar guides for computers: Rules that tell the computer how verbs change (like "I go," "he goes"), how words are ordered, or how a word's meaning changes based on other words around it.

The issue is, we don't have good, comprehensive digital dictionaries or detailed, computer-friendly grammar rules specifically for Rajasthani Language. It's like trying to teach a language when you don't have a full dictionary or a clear grammar book to rely on. The computer struggles to understand individual Rajasthani Language words or even build a correct Rajasthani Language sentence from scratch.

c) No Good Test Scores or Ways to Compare

Once you try to build a computer translator, how do you know if it's any good? How do you compare it to another translator?

For major languages, there are standard "test scores" (called "evaluation metrics") and "example tests" (called "benchmarks"). These help you objectively measure how accurate and natural-sounding a translation is. For example, if you translate a sentence and a human expert says it's perfect, the computer gets a high score.

But for Hindi-to-Rajasthani Language, because so little research has been done, we don't have these agreed-upon "test scores" or standard "example tests." It's like trying to grade a student's performance without a clear grading system or knowing what a "good" answer looks like. This makes it very hard to tell if a new translation method is actually an improvement or just different.

d) Not Enough Computer-Friendly Language Research

Language experts have studied Rajasthani Language, but usually not in a way that helps computers directly. "Computational Linguistic Analysis" means studying a language specifically to understand how computers can process and translate it.

This involves deep dives into things like:

- How exactly Rajasthani Language words change their endings (like "go," "going," "went").
- How sentences are structured differently than in Hindi.
- How to handle unique Rajasthani Language phrases or cultural sayings.

Currently, there hasn't been enough of this specific type of research for Rajasthani Language. We don't have enough detailed blueprints that tell a computer precisely how to handle Rajasthani Language's unique quirks when translating.

e) Untried Teaching Methods for Rajasthani Language

There are different "teaching methods" or "models" that computers use to translate languages. Some are older, like "Statistical Machine Translation" (which counts how often words appear together). Newer ones are "Neural Machine Translation" (which use complex computer networks similar to simplified brains).

The problem is, the methods that work really well for languages with tons of data (like English or Hindi) often don't work so well, or haven't been properly adjusted, for languages like Rajasthani Language that have very little data.

So, we haven't really tested which "teaching method" is the best fit for Hindi-to-Rajasthani Language, given how little data we have. We also haven't figured out how to tweak or "optimize" these methods to get the best results specifically for Rajasthani Language's unique features. It's like having powerful teaching tools but not knowing which one is best for a student who has very few study materials.

This study seeks to address the problem by creating a Hindi to Rajasthani Language machine translation system using NLP techniques. It aims to identify how a translation system can be developed when standard linguistic tools and data are missing. This involves the creation of a

parallel dataset, linguistic analysis, selection and training of translation models, and evaluation.

This research project is all about creating a good, easy-to-use, and accurate machine translation system for Hindi to Rajasthani Language. The main problem we're tackling is that there aren't enough high-quality language resources (like paired sentences in both languages or digital dictionaries and grammar rules) and ready-made computer models specifically designed for translating between these two languages, especially since Rajasthani Language is considered a "low-resource" language in the digital world.

Key issues that will be explored include:

1. How to collect and structure Hindi-Rajasthani Language parallel data in the absence of a standardized script and grammar?
2. What translation techniques (rule-based, statistical, or neural) are most suitable for such low-resource scenarios?
3. How can the quality of translation be evaluated for a dialect with no benchmark or gold standard?
4. Can multilingual pre-trained models or transfer learning approaches be fine-tuned for Hindi-Rajasthani Language translation?

3.2 Significance / Rationale of the Study

The significance of this study lies in its potential to make an important contribution to both language technology and cultural preservation. Firstly, Rajasthani Language, like many other Indian dialects, is at risk of being lost. As formal education and digital content are largely available in Hindi and English, young Rajasthani Language speakers are losing touch with their mother tongue. This study can support the preservation and revival of Rajasthani Language by creating computational tools that allow its use in digital formats.

Secondly, this research addresses a real communication gap. Many Rajasthani Language speakers in rural regions face difficulty understanding government notifications, legal documents, health advisories, and educational material published in Hindi. A translation system from Hindi to Rajasthani Language will bridge this language gap and make important information accessible in a familiar language.

Thirdly, this research will be the first of its kind to attempt computational modeling for Rajasthani Language. Since no machine translation work has been done on this language pair, the findings of this study will be original and valuable. It will serve as a reference for future scholars and researchers interested in other low-resource or under-documented Indian languages.

Fourth, the study will contribute to the larger goal of inclusive digital development in India. Government initiatives like Digital India and Bhashini aim to make technology available in all Indian languages. By focusing on Rajasthani Language, this research supports the inclusion of marginalized dialects into the national digital ecosystem.

This study is also motivated by the availability of modern NLP technologies that can be adapted for low-resource settings. Models like mBART, IndicTrans, and other transformer-based architectures have shown promise for training on small datasets using transfer learning and multilingual training. This makes it both timely and technically feasible to work on Rajasthani Language translation.

This research project is all about creating a good, easy-to-use, and accurate machine translation system for Hindi to Rajasthani Language. The main problem we're tackling is that there aren't enough high-quality language resources (like paired sentences in both languages or digital dictionaries and grammar rules) and ready-made computer models specifically designed for translating between these two languages, especially since Rajasthani Language is considered a "low-resource" language in the digital world.

3.3 Research Questions

1. How can a high-quality parallel dataset for Hindi-Rajasthani Language be built from scratch?
2. Can modern multilingual models be successfully fine-tuned for Rajasthani Language despite its lack of astandardized script or grammar?
3. Which translation technique is the most effective for this low-resource language pair?
4. What are the societal and educational benefits of Hindi-Rajasthani Language translation tools?

3.4 Objectives of the Study

1. To build a Hindi-Rajasthani Language parallel dataset through data collection and digitization.
2. To develop a hybrid machine translation system optimized for low-resource settings.
3. To evaluate translation quality using both automated metrics and human assessment.

4. Methodology

4.1 Research Design

The methodology of research focuses on the design, development, and evaluation of a machine translation system that converts Hindi text into Rajasthani Language using Natural Language Processing (NLP) techniques. Due to the low-resource nature of Rajasthani Language and the absence of formal corpora or digital tools, a hybrid approach combining data-driven modelswith rule-based linguistic features will be used. Our research will use a combination of designs:

Evaluative: This means we will build a translation system and then carefully measure how well it works. We will check its accuracy and how natural its translations sound.

Exploratory: Because translating low-resource languages like Rajasthani Language is a tough problem with limited existing solutions, we will be exploring and trying out new ways and combinations of advanced computer techniques to find the best approach.

Comparative: We will compare different translation methods and models to see which ones give the best results for Hindi to Rajasthani Language.

Design: Strategy & Framework

Step 1: Data Collection

Collect Hindi texts and translate them into Rajasthani Language to build a dataset (called a "parallel dataset"). Collect Rajasthani Language texts from books, folk stories, and local speakers.

Step 2: Data Processing

Clean the texts by removing unwanted symbols, correcting errors, and aligning sentences. Tag parts of speech (like noun, verb, adjective) and handle differences between dialects.

Step 3: Model Development

Rule-Based: Build rules to translate Hindi words and grammar into Rajasthani Language.

Statistical: Use past data to find patterns in translation.

Neural (AI-based): Use machine learning models like Transformers for better accuracy.

Step 4: Testing and Evaluation

Test the translation system using software tools (like BLEU score).

Also ask real people who know Hindi and Rajasthani Language to check how good the translations are.

4.2 Sample/Participants

A **non-probability sampling strategy** is almost always more appropriate and practical than probability sampling. This is because you are not trying to generalize findings about a population of people, but rather about the linguistic phenomena themselves.

Research on Hindi to Rajasthani Language translation, a **non-probability sampling** strategy will be used. Specifically, **purposive sampling** will be applied. This means the samples (data/texts) will be **selected based on specific purposes** of the study—such as representing common and meaningful language usage patterns in both Hindi and Rajasthani Language.

Since the research focuses on regional language translation where standardized datasets are not available, it is not possible to use random sampling methods. However, the **quality and representativeness of the selected texts and participants** will be ensured.

Inclusion / Exclusion Criteria

Inclusion Criteria:

Texts and documents written in **standard Hindi** and **authentic Rajasthani Language dialects** (spoken and written).

News articles, folk stories, religious texts, local government documents

Participants : Native Rajasthani Language speakers who are fluent in Hindi and can help in human evaluation of the translations.

Exclusion Criteria:

Texts that are **highly technical or domain-specific** (e.g., medical, legal) as they may not reflect general language usage.

Non-native speakers or individuals with limited knowledge of either Hindi or Rajasthani Language (if human validation is part of the study).

Sample Recruitment and Approach

The sample for the study will be gathered from:

Online sources like blogs, news websites (e.g., Dainik Bhaskar, Rajasthan Patrika), and digital archives of Rajasthani Language literature.

Printed texts like books and folk stories will be digitized using OCR tools.

For validation or evaluation, **language experts and native speakers** will be approached

Collaborations: Reach out to local linguistic experts, historians, or cultural organizations in Rajasthan who might have access to relevant historical texts, folk literature, or translated works.

Literary Works: Access published Hindi literature and seek out their Rajasthani Language translations (if available).

Sample Size and Determination

The sample size will be determined based on:

dataset requirement for training NLP models (like NMT – Neural Machine Translation)

Minimum of **20,000 to 50,000 aligned sentence pairs** (Hindi-Rajasthani Language) for initial model training and testing

For human evaluation, 20 to 30 bilingual participants will be selected to evaluate the translation quality using standard metrics like BLEU score, adequacy, and fluency

This is a **low-resource language pair**, so even a smaller, high-quality dataset can significantly contribute.

4.3 Operational Definitions of the Variables

a. Source Language (SL): Hindi

Construct Definition: For the purpose of this study, "Source Language (SL): Hindi" refers to the standard variety of Modern Standard Hindi (MSH) as predominantly used in contemporary written texts (e.g., news articles, literary prose, official documents) and educated conversational contexts.

b. Target Language (TL): Rajasthani Language

Construct Definition: "Target Language (TL): Rajasthani Language" refers to a specific dialectal variety of Rajasthani Language, primarily focusing on the [Specify the dialect, e.g., Jodhpuri Rajasthani Language, Bikaneri Rajasthani Language, or a representative standard if one is emerging] spoken in [Specify geographic area, e.g., the Jodhpur region of Rajasthan]. This study acknowledges the internal variation within Rajasthani Language but will primarily focus on features common to or representative of the chosen dialect.

c. Translation Process / Translation Strategies

Construct Definition: "Translation Process / Translation Strategies" refers to the observable techniques and methods employed to render meaning from Hindi to Rajasthani Language, addressing linguistic, cultural, and pragmatic challenges. This construct is not about the cognitive process *of* the translator, but rather the *results* of the application of specific strategies evident in the translated text.

Data Collection and Parallel dataset : The first and most critical step is to gather a substantial amount of high-quality parallel data (Hindi-Rajasthani Language sentence pairs). Since a ready-made, large-scale dataset is likely unavailable, we would need to build one with, Manual translation of Hindi texts into Rajasthani Language, Leveraging any existing texts (e.g., from literature, folklore, government documents)

Data Cleaning: collected data needs to be pre-processed using Tokenization (breaking down text into words or sub-words), Normalization (converting text to a standard format, e.g., handling different spellings or transliterations), Removing special characters and noise.

Model Architecture and Transformer model **:** use, **Encoder** for input Hindi sentence. that creates a contextualized representation of each word, taking into account all the other words in the sentence. **Decoder** takes the encoder's output and generates the translated Rajasthani Language sentence, one word at a time. It also uses an attention mechanism to focus on relevant parts of the input Hindi sentence.

Transfer Learning: We can start with a pre-trained model on a related, high-resource language pair (e.g., Hindi-English) and then fine-tune it with your limited Hindi-Rajasthani Language parallel data. This allows the model to leverage the knowledge it has already gained about language structure and translation.

Training : The model is trained on the parallel dataset to minimize the difference between its predicted output and the actual target translation.

Evaluation: The performance of the translation system is typically measured using metrics such as BLEU (Bilingual Evaluation Understudy) score, which measures the similarity of the machine's translation to a human-translated reference.

d. Translation Challenges/Difficulties

Construct Definition: "Translation Challenges/Difficulties" refers to specific linguistic, cultural, or pragmatic obstacles encountered during the translation process from Hindi to Rajasthani Language, which may lead to shifts, errors, or a perceived loss of meaning/nuance.

e. Translation Quality

Construct Definition: "Translation Quality" refers to the effectiveness and appropriateness of the Rajasthani Language translation in accurately, fluently, and naturally conveying the meaning and intent of the Hindi source text, while adhering to the linguistic and cultural norms of the target audience.

Important of operational definitions in research:

Consistency: Ensures you are identifying and analyzing the same phenomena throughout your entire research.

Clarity: Makes your methodology transparent to your readers, allowing them to understand exactly what you are measuring.

Foundation for Analysis: Provides the categories and criteria you will use for your data collection, annotation, and subsequent analysis.

4.4 Data Collection:

a) Secondary Data Collection

Secondary data" refers to data that has already been collected, processed, and often published by others. This is often the backbone of linguistic dataset-based research.

Type of Secondary Data:

Books and Literature, Online Resources, Digital Libraries and Corpora, Government Documents.

Existing Corpora: Hindi Corpora, Rajasthani Language Corpora, Parallel Corpora, Comparable Corpora, Dictionaries and Grammars, Existing Translation Studies/Theories, Literary Translations, Historical Documents

b) Data Collection Procedure

1: Initial dataset Building (Secondary Data Focus): Begin by systematically identifying, acquiring, and pre-processing existing Hindi and Rajasthani Language texts, especially parallel corpora.

2: Gap Identification & Primary Data Need Assessment: During initial review of secondary data, identify specific linguistic phenomena, translation challenges, or cultural nuances that are not adequately represented or explained. This will inform the need for primary data.

3: Primary Data Collection (if required): If gaps are identified, execute the primary data collection plan (e.g., elicitation tasks with native speakers, expert interviews, collection of specific contemporary texts).

4: Data Integration and Final dataset Formation: Combine all collected primary and secondary data into a cohesive and organized dataset. This might involve normalization, consistent formatting, and further annotation.

5: Data Management and Storage: Outline how the data will be organized, stored securely (e.g., cloud storage, external hard drives), and backed up to prevent loss. Mention data anonymization if human participant data is involved.

c) Ethical Considerations

Consent: Written or verbal consent will be taken from all native speakers involved in data contribution (interviews, translations, etc.).

Access: Only publicly available or permitted data will be used. No copyrighted content will be included without proper permission.

Participants' Protection: All participant identities will remain **anonymous**.

Data collected from individuals will be used **strictly for academic purposes**.

All recordings, texts, or responses will be stored securely and not shared publicly without consent.

5. Limitations

1. Small amount of data: There are not many Hindi–Rajasthani Language sentence pairs available for training.
2. Many Different Dialects: The current model was trained on a specific dialect of Rajasthani Language, and its performance may vary when translating other regional variants.
3. Trouble with Names and New Words: Integrating a transliteration module could further enhance the translation of proper nouns and out-of-vocabulary words.

4. Future research could explore more advanced architectural approaches, such as transformer-based models with pre-trained multilingual embeddings, to push the performance boundaries even further.

6. Work Plan of the Research

Phase 1: Project Literature Review & Problem Definition (Months 1–4)

Objective Definition, Technology Stack Selection.

Phase 2: Data Collection & dataset Development (Months 5–10)

Rajasthani Language Data Translation, Data Annotation & Quality Control.

Phase 3: Model Design & Selection (Months 11-15)

Text Cleaning, Tokenization & Lemmatization, Vocabulary Creation & Embeddings.

Phase 4: Model Development & Training, Code Optimization (Months 16–24)

Model Architecture Design, Model Training.

Phase 5: Evaluation & Analysis (Months 25–28)

Automated Evaluation, Human Evaluation, Error Analysis & Fine-tuning.

Phase 6: Documentation & Thesis Writing (Months 28-36)

Deployment, User Interface Development, Feedback Mechanism,

Maintenance & Updates

7. References

[1]. Khadka, P., Bk, A., Acharya, A., K.C., B., Shrestha, S., & Thapa, R. (2025, January). Nepali Transformers @ NLU of Devanagari Script Languages 2025: Detection of Language, Hate Speech and Targets.

[2]. Rathod, D., Yadav, A. K., & Kumar, M. (2025, February 18). Character-level encoding based neural machine translation for Hindi language. Journal Name, 57(23).

[3]. Kumar, A., & Goyal, V. (2018, July). Hindi to Punjabi machine translation system based on statistical approach. Journal of Statistics and Management System. https://doi.org/10.1080/09720510.2018.1466963

[4]. Yadav, S., Prasad, S. S., Maurya, S., Garg, R., Kashikar, A., & Shelke, N. (2025, January 16–18). Multilingual Sentiment Analysis: A Comprehensive Review for Indian Languages. IEEE.

[5]. Anwaya, K., Ananya, C., & Balipa, M. (2025, April 24–25). Natural Language Processing Model for Hindi to Tulu Translation. IEEE.

[6]. Chouhan, S., Nath, S. B., & Dutta, A. (2024, December 3). HindiLLM: Large Language Model for Hindi. Journal Name, pp. 255–270.

[7]. Agarwal, S., Raj, L. A., & Dutta, S. (2025, January 18–19). Improvising language conversation – Hindi & Urdu. IEEE. Electronic ISSN: 2688-0288.

[8]. Mohapatra, U. M. (2025). Machine learning-based translation from Odia to Hindi. IEEE. ISSN: 25836404.

[9]. Mete, A. M., & Dhore, M. L. (2025). Abstractive text summarization for Hindi language using T5 Transformer. IEEE. ISSN: 25836404.

[10]. Harish, B. S., & Rangan, R. J. (2020, June 12). A comprehensive survey on Indian regional language processing. IEEE. ISSN: 25836404.

[11]. Simiscuka, A. A. (2023, May 19). Research on negative post identification in the regional language (Hindi). IEEE. ISSN: 25836404.

[12]. Santhi, J. A., Bhavatharini, S., & Nandhini, G. G. (2022, July 26). Study of regional language translator using natural language processing. Vol. 302, pp. 91–100.

[13]. Desai, N. P., & Dabhi, V. K. (2021, January 30). Taxonomic survey of Hindi language NLP system.

[14]. Sethi, N., Dev, A., & Bansal, P. (2023, January 8). A novel neural machine translation approach for low-resource Sanskrit-Hindi language pair.

[15]. Sirsat, S., & Zulpe, N. (2023, March 23). Efficient dataset preparation techniques for regional/Marathi language analysis: Creating customized dataset. IEEE. ISBN: 979-8-3503-3330-5.

[16]. Sethi, N., Dev, A., & Bansal, P. (2023, December 15). Enhancing low-resource Sanskrit-Hindi translation through deep learning with Ayurvedic text. IEEE.

[17]. Jha, A., Hemprasad, & Patil, Y. (2022, November 25). A review on machine transliteration, translation, evaluation metrics and datasets in Indian languages. IEEE, 82.

[18]. Kazi, M., Mehta, H., & Bhari, S. (2020, December 17). Sentence level language identification in Gujarati-Hindi code mixed script. IEEE. ISBN: 978-1-7281-8881-2.

[19]. Beshears, A. (2017, October 13). The demonstrative nature of the Hindi/Rajasthani Language correlative. Queen Mary University of London. ISBN: 978-1-7281-8881-2.

[20]. Joshi, R., Karnavat, R., Jirapure, K., & Joshi, R. (n.d.). Evaluation of deep learning models for hostility detection in Hindi text.

[21]. Sadaram, V. S., & Badugu, S. (n.d.). Interactive Telugu-language text dialogue system using keyword extraction and semantic parsing.

[22]. Chikkarangaiah, J., Uday, A., De Hilster, D., Gangadhar, S., & Shetty, J. (2024, September). Enhancing the English natural language processing dictionary using NLP++. IAES International Journal of Artificial Intelligence (IJ-AI), 13(3), 3466–3477. https://doi.org/10.11591/ijai.v13.i3.pp3466-3477

[23]. Rabha, S., Agarwal, R., Sarthak, S., Parikh, C., & Sharma, A. (n.d.). Recent trends in speech recognition for diverse Indian languages.

[24]. Shah, P., & Bakrola, V. (2020). Attention-based Encoder–Decoder for Hindi–Gujarati Translation. In

Proceedings of the 1st International Conference on Advances in Computing and Communication Technology (ICACCT).

[25]. Goyal, V., Kumar, S., & Sharma, D. M. (2020). Multilingual Transfer & Transliteration for Low-resource Indo-Aryan NMT. In

Proceedings of the 17th International Conference on Natural Language Processing (ICON).

[26]. Dewangan, S., Alva, S., Joshi, N., & Bhattacharyya, P. (2021). Benchmarking Indic NMT with Subword and BERT-augmented Methods.

ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 21(3), 1–25.

[27]. Kumar, A., Parida, S., Pratap, A., & Singh, A. K. (2023). Phonetic-orthographic Encoding for Low-resource Indic NMT.

[28]. Das, S. B., Panda, D., Mishra, T. K., Patra, B. K., & Ekbal, A. (2023). Indic→Indic MNMT Baselines on Samanantar. In

Proceedings of the 23rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing).

[29]. Roy, M. K., Arora, K. K., & Arora, S. (2023). MTIL-FIRE2023: Bidirectional Hindi–Punjabi NMT with NLLB-200. In

Proceedings of the Forum for Information Retrieval Evaluation (FIRE).

[30]. Dewangan, S., Alva, S., Joshi, N., & Bhattacharyya, P. (2021). ILNMT: Large Empirical Study of Neural MT for Indian Languages.

[31]. Journal of Artificial Intelligence Research, 71, 1029–1068.

[32]. Das, S. B., Panda, D., Mishra, T. K., Patra, B. K., & Ekbal, A. (2023). Indic→Indic MNMT Baselines on Samanantar. In

Proceedings of the Forum for Information Retrieval Evaluation (FIRE).

[33]. Laskar, S. R., Manna, R., Pakray, P., & Bandyopadhyay, S. (2022). Transliteration-based Multilingual NMT for Indic Languages (WAT2022). In

Proceedings of the 13th Workshop on Asian Translation (WAT).

[34]. Authors — Sentiment-Aware Machine Translation for Indic Languages (2025). Sentiment-Aware Machine Translation for Indic Languages. In

[Unspecified Journal/Conference Proceedings].

[35]. Madaan, L., Sharma, S., & Singla, P. (2020). Transfer-Learning Setups for Similar Language Translation at WMT20. In

Proceedings of the Fifth Conference on Machine Translation (WMT).

[36]. Kumar, A., Baruah, R., Mundotiya, R. K., & Singh, A. K. (2020). Transformer-based Neural Machine Translation System for Hindi–Marathi. In

Proceedings of the Fifth Conference on Machine Translation (WMT).

[37]. Laskar, S. R., Khilji, A. F. U. R., Pakray, P., & Bandyopadhyay, S. (2020). Hindi–Marathi Cross-Lingual Model (Team CNLP-NITS). In

Proceedings of the Fifth Conference on Machine Translation (WMT).

[38]. Desai, S., Kshirsagar, A., & Marathe, M. (2021). Multitask Finetuning for Improving Neural Machine Translation in Indian Languages. In

Proceedings of the 18th International Conference on Natural Language Processing (ICON).

[39]. Bollegala, D. et al. (2021). Bridging Lexical Divergence for Unsupervised Hindi–Marathi MT.

[40]. Ala, H., & Sharma, D. M. (2020). AdapNMT: Neural Machine Translation with Technical Domain Adaptation for Indic Languages.

[41]. Ala, H., Mujadia, V., & Sharma, D. M. (2021). Domain Adaptation for Hindi–Telugu MT using Domain-Specific Back-Translation. In

Proceedings of the 1st Workshop on Language Technology for Global Health.

[42]. Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., Raghavan, A. K., Sharma, A., Sahoo, S., et al. (2021). Samanantar: The Largest Public Parallel Corpora Collection for 11 Indic Languages. In
Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.

[43]. Mujadia, V., & Sharma, D. M. (2022). The LTRC Hindi–Telugu Parallel Corpus. In Proceedings of the 2nd Workshop on Indian Language Technology (ILT).